Algerian Democratic and Popular Republic

Ministry of Higher Education and Scientific Research

University of August 20, 1955 - Skikda

Faculty of Sciences

Department of Agronomic sciences

Biostatistics course

According to the official educational program of third year Agroecology semester 3

Developed by

Dr. BELAIDI Abdelouahab

Lecturer class B

Department of Agronomic sciences

Academic year: 2024/2025

Preamble

The aim of teaching biostatistics is to provide second-year students with essential skills and knowledge to analyze, interpret, and draw conclusions from biological data. The objectives of teaching biostatistics are:

- Understanding of fundamental statistical concepts: students should gain a thorough understanding of basic statistical concepts such as probability, distribution, mean, median, standard deviation, etc.
- Data collection and analysis: students should learn data collection techniques, including the design of experimental studies, sample selection and collection of relevant data. They should also be able to analyze these data using appropriate statistical methods.
- Interpretation of results: Instruction should enable students to correctly interpret statistical results. This includes the ability to draw meaningful conclusions from statistical analyses, assess the validity of results, and recognize study limitations.

To achieve these goals, the student must have notions of probability and numerical analysis already seen in the subject Mathematics in the first year.

Table of content

Chapter 1: One and two dimensional descriptive statistics

Chapter 2: Statistical methods relating to means

Chapter 3: Statistical methods relating to dispersion

Chapter 01: One and two dimensional descriptive statistics

Introduction

Descriptive statistics is a fundamental branch of statistics that focuses on summarizing and interpreting data to reveal patterns and essential characteristics. It involves collecting, organizing, presenting, and analyzing data in a way that provides a clear and concise understanding of the information at hand. Descriptive statistics are widely used in various fields such as economics, biology, social sciences, and engineering to make data-driven decisions and communicate findings effectively. In this chapter, we will explore **descriptive statistics with one and two dimensions**, examining methods and techniques used to describe data distributions and relationships.

1- One-dimensional descriptive statistics

1.1- Objective

The objective of **one-dimensional descriptive statistics** is to summarize and describe the main features of a single variable within a dataset.

1.2- Statistical character

1.2.1- Definition

A characteristic refers to a quantity or attribute that can be observed in an individual and is capable of varying, thus taking on different states called **modalities**.

1.2.2- Types of statistical characters

1.2.2.1- Qualitative character

Measured on a nominal scale, the modalities are expressed by names and are not hierarchical.

Examples: color, blood groups, different nucleotides in DNA, the presence or absence of a characteristic, etc.

1.2.2.2- Quantitative characters

A character is **discrete** if it can only take certain values within a given interval. Generally, it results from counting or enumeration.

Examples: the number of offspring per litter, the number of cells in a culture, the number of accidents over a given period, etc.

A character is **continuous,** if it can theoretically take any value within a given interval. Generally, it results from measurement.

Examples: weight, height, blood glucose level, yield, etc.

1.3- Data representation

There are several levels of statistical description: raw data presentation, numerical table presentations, graphical representations, and numerical summaries provided by a small number of characteristic parameters.

1.3.1- Statistical series

A statistical series corresponds to the different modalities of a characteristic in a sample of individuals belonging to a given population. The number of individuals that make up the studied sample is called the **sample size**.

1.3.2- Statistical table

The frequency distribution table is a synthetic way of presenting data. Its construction is immediate in the case of a discrete character but requires data transformation in the case of a continuous character.

Discrete quantitative character: In the case of a discrete quantitative characteristic, the establishment of the distribution of observed data along with their frequencies is immediate.

Continuous quantitative character: In the case of a continuous quantitative characteristic, the construction of the frequency table requires prior grouping of the data into classes. This involves defining the expected number of classes and thus the width of each class or class interval.

- Various empirical formulas allow for determining the number of classes for a sample of size *n*.

Sturges' rule: Number of classes = 1+3.3log n.

- The class interval is then obtained as follows: **Class interval = max-min / Number of classes**
Where;

X_{max} and X_{min} are the largest and smallest values of X in the statistical series, respectively.
Starting from X_{min}, the class limits or boundaries are obtained by successively adding the class interval. Generally, an attempt is made to align the class index or central value of the class with a whole number or one with few decimals.

1.3.3- Graphical representations

Graphical representations have the advantage of providing immediate insight into the general shape of the distribution. They facilitate the interpretation of collected data.

1.3.3.1- Discrete quantitative character

For discrete quantitative character, the graphical representation is a **bar chart**, where the height of the bars corresponds to the frequency n_i associated with each modality X_i of the character.

1.3.3.2- Continuous quantitative character

For continuous quantitative character, the graphical representation is a **histogram**, where the height of the rectangle is proportional to the frequency *ni*. This is only true if the class interval is constant. In this case, the area under the histogram is proportional to the total frequency.

1.3.4- Numerical indicators

The final level of statistical description is the numerical summary of a statistical distribution using numerical indicators or characteristic parameters.

1.3.4.1- Position indicators

These parameters aim to characterize the order of magnitude of observations in the case of a quantitative character.

1.3.4.1.1- Arithmetic mean ($\bar{x}$)

For a sample of n observed values $x_1, x_2,\ldots,x_i,\ldots,x_n$ of a quantitative character X, the observed mean is defined as the arithmetic mean of the n values.

$$\boxed{\bar{X}=\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Remark: One of the properties of the arithmetic mean is that the sum of deviations from the mean is zero. If the observed data x_i are grouped with frequency n_i , they must be weighted by the corresponding frequencies.

$$\boldsymbol{\overline{X}= \frac{1}{n}\sum_{i=1}^{n} x_i n_i}$$

1.3.4.1.2- Median (Me)

The median *Me* is the value of the characteristic for which the cumulative frequency is equal to 0.5 or 50%. It corresponds to the center of the statistical series arranged in ascending order, or the value for which 50% of the observed values are higher and 50% are lower.

- If the values taken by the characteristic are not grouped into classes:
  - If *n* is odd, the median is the middle value $X_{(n+1)/2}$
  - If *n* is even, the median is any value between $\mathbf{X_{\frac{n}{2}}}$ **et** $\mathbf{X_{\frac{n}{2}+1}}$

In this case, it may be convenient to take the midpoint.

- If the values taken by the characteristic are grouped into classes, the class containing the $n/2^{th}$ individual of the sample (median class) is identified. The exact position of the $n/2^{th}$ individual is determined by linear interpolation:

$$\boldsymbol{m_e = x_m + (i)\left(\frac{\frac{n}{2} - N_i}{n_i}\right)}$$

Where:

- ***i:*** class interval ;
- x_m **:** lower limit of the class the median class;
- n_i**:** frequency of the median class;
- ***Ni*:** cumulative frequency below *xm*;
- ***n*:** sample size.

1.3.4.1.3- Mode *(Mo)*

Let X be a character of a sample of size n, the mode is a value of X of maximum frequency. If the values taken by the characteristic are grouped into classes, the mode corresponds to the class with the maximum frequency in the frequency distribution. The mode can be identified as the median value of the class with the maximum frequency or determined by linear interpolation as follows:

$$\mathbf{M_o} = x_m + \frac{\Delta i \times i}{\Delta x + \Delta i}$$

- x_m**:** lower limit of the class with the maximum frequency;
- ***i*****:** class interval (x_{m+1}–x_m);
- **Δ*i*:** frequency difference between the modal class and the nearest lower class;
- **Δ*s*:** frequency difference between the modal class and the nearest higher class.

2- Two-dimensional descriptive statistics

2.1- Objective

The objective of two-dimensional descriptive statistics is to analyze and summarize the relationship between two variables, allowing us to understand how they interact or correlate with each other. It helps to identify patterns, associations, or dependencies between the variables and quantify the strength and direction of their relationship.

2.2- Presentation of a bivariate data

Bivariate statistical series can be presented in these ways.

2.2.1- Bivariate statistical series

A bivariate statistical series X and Y is any list of pairs of the type: (x_1, y_1), (x_2, y_2), ...(x_n, y_n), where n is the number of pairs. The data is arranged as shown in the following table:

Observation	1	2	……..	n
Variable X	X_1	X_2	………	X_n
Variable Y	Y_1	Y_2	………	Y_n

2.2.2- Scatterplots

Bivariate data are usually represented graphically on scatterplots.

A scatter plot is a graph that shows whether there is a relationship between two variables x and y. Each data value (xi , yi) on a scatter plot is shown by a point on a Cartesian plane.

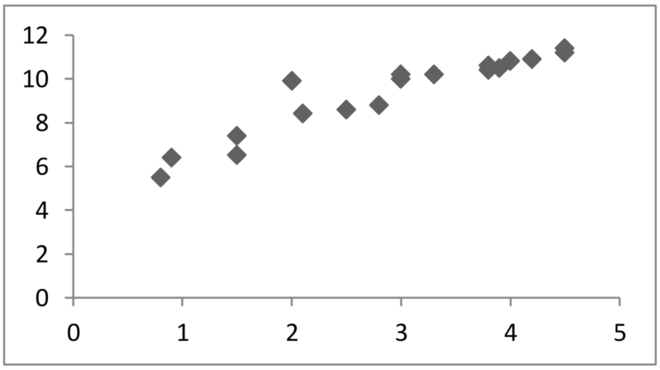

Figure 1. Scatterplots

2.2.3- Contingency table

A contingency table is a table that displays data for one variable in rows and data for another variable in columns.

Y \ X	X_1	X_2	……..	X_i
Y_1	n_{11}	n_{21}	………	n_{i1}
Y_2	n_{12}	n_{22}	………	n_{i2}
Y_j	n_{1j}	n_{2j}	……...	n_{ij}

n_{ij} : frequency of the pair (xi, yj).

2.2.3.1- Marginal frequency:

- Marginal frequency of xi **($n_{i\bullet}$)** is given by:

$n_{i\bullet}$ = X l j =1 $n_{ij} = n_{i1} + n_{i2} + ... + n_{il}$ = total de la ligne i

- Marginal frequencye of yj **($n_{\bullet j}$)** is given by:

$n_{\bullet j}$ = X k I =1 $n_{ij} = n_{1j} + n_{2j} + ... + n_{kj}$ = total de la colonne j

2.2.4- Covariance

The covariance is a measure of relationship between two variables. Let X and Y be two variables then covariance between them is given by:

$$\boldsymbol{cov\ (x, y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}$$

or

$$\boldsymbol{cov\ (x, y)} = \frac{1}{n}\left[\sum x_i y_i - \frac{\sum x_i \times \sum y_i}{n}\right]$$

2.2.4.1- Interprétation:

Covariance indicates whether the variables x and y vary in the same direction or in two opposite directions:

- If cov(x, y) > 0 then the two variables are not independent, they are positively related and vary in the same direction.

- If cov(x, y) < 0 then the two variables are not independent, they are negatively related and vary in two opposite directions.

- If cov(x, y) = 0 then variations in one of the two variables do not cause variations in the other.

2.3- Simple linear regression

It aims to establish a link between a dependent variable Y and an independent variable X to then be able to make predictions about Y when X is measured.

The connection between the dependent variable Y and the independent variable X can be modeled by a function of type Y = ax + b, represented graphically by a straight line.

2.3.1- Estimation of the equation of the regression line (Y = ax + b)

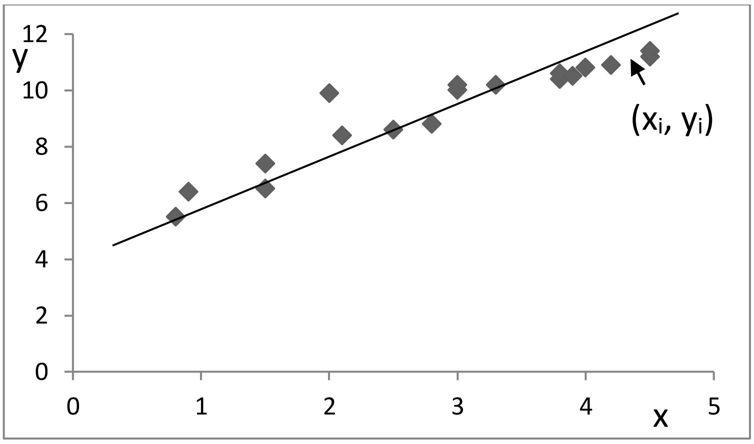

Figure 2. Regression line

We speak of linear regression when we wish to calculate a 1st degree function linking the variables x and y, this linear function of the form: y=ax+b

Y: dependent variable (explained)

X: independent (explanatory) variable

a: regression coefficient

b: value of Y for x = 0

2.3.1.1- Calculation of the regression coefficient (a)

The regression coefficient can be calculated by applying the least squares method, we obtain:

$$\boxed{a= \frac{cov(x,y)}{S_x^2} = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum(x_i-\bar{x})^2} = \frac{\frac{1}{n}\left[\sum x_i y_i - \frac{\sum x_i \times \sum y_i}{n}\right]}{\frac{1}{n}\sum x_i^2 - \bar{x}^2}}$$

2.3.1.2- Calculation of (b)

A particularity of the regression line is that it passes through the theoretical average point with coordinate:

therefore : $\bar{y}$ = a$\bar{x}$+b

$$\boxed{b= \bar{y} - a\bar{x}}$$

2.4- Linear correlation

Linear correlation is a measure of dependence between two random variables. The most commonly used coefficient is the linear correlation coefficient of Bravais Pearson (r).

2.4.1-Linear correlation coefficient

The linear correlation coefficient measures the strength and the direction of the linear relationship between two variables x and y.

$$\boxed{\mathbf{r} = \frac{\boldsymbol{cov}\ (\boldsymbol{x},\boldsymbol{y})}{\sqrt{\boldsymbol{S_x^2} \times \boldsymbol{S_y^2}}}}$$

2.4.1.1- Interpretation

The correlation coefficient r ranges between [-1; 1].

- **r = -1 or r = 1:** there is perfect correlation between X and Y and the points (Xi; Yi) are all on the regression line.
- **r = 0**: there is no correlation between X and Y and the points (Xi; Yi) are scattered randomly.
- **0 < r < 1:** then there is a weak, medium or strong positive (increasing) correlation between X and Y. In this case, an increase in X leads to an increase in Y.
- **1 < r < 0:** then there is a weak, medium or strong negative (decreasing) correlation between X and Y. In this case, an increase in X leads to a decrease in Y.

Chapter 2: Statistical methods relating to means

Introduction

The analysis of means plays a crucial role in inferential statistics. It allows us to evaluate whether an observed mean conforms to a reference value (**conformity test**) or whether two means differ significantly (**comparison test**). These tests are often based on the normal distribution or Student's t-distribution, depending on sample sizes and the knowledge of variances.

1- Conformity test and confidence interval of a mean

The purpose of this test is to determine whether the mean μ of a population is equal to a theoretical value or standard (μ_0). We have a sample of size n, with a mean and standard deviation S drawn from a population with mean μ and standard deviation σ. The question is whether the population mean is equal to a given value (μ_0).

We can answer this question by either verifying if μ_0 is within the confidence interval of $\bar{x}$ or by performing a conformity test. This test includes the following steps:

1.1- Conditions of application

The use of the test is subject to the following conditions:

- The data follows a normal distribution;
- The sample are random.

1.2- Test Hypotheses

- **Null hypothesis (H_0): $\mu = \mu_0$ (**the population mean μ is equal to μ_0).
- **Alternative hypothesis (H_1): $\mu \neq \mu_0$** (the population mean μ is different from μ_0).

1.3- Test procedure

1.3.1 - Population variance (σ^2) is unknown

1.3.1.1 - Sample size (n) < 30

The test is performed by calculating the quantity $\mathbf{t_{observed}}$ and drawing $\mathbf{t_{critical}}$.

- $\mathbf{t_{obs}} = \frac{|\bar{x} - \mu 0|}{\frac{S}{\sqrt{n}}}$

Where:

$\bar{x}$ sample mean, μ_0 theoretical value, and S sample standard deviation.

- **Critical value:** $t_{1-\alpha/2}$ from Student's t distribution with (n -1) degrees of freedom.

Confidence interval

$$CI = \bar{x} \pm [\ t_{1-\alpha/2} \times \frac{s}{\sqrt{n}}]$$

1.3.1.2 - Sample size (n) ≥ 30

To test the null hypothesis ($\mu = \mu_0$) against the alternative hypothesis ($\mu \neq \mu_0$) when we have a large sample, we form the quantity:

- $\mathbf{Z_{obs}} = \frac{|\bar{x} - m_0|}{\frac{S}{\sqrt{n}}}$
- **Critical value**: $Z_{1-\alpha/2}$ from the standard normal distribution Z.

Confidence interval

$$CI = \bar{x} \pm [\ \mu_{1-\alpha/2} \times \frac{s}{\sqrt{n}}]$$

1.3.1.3 - Population variance (σ^2) is known

If the population variance is known, μ follows the normal distribution

- $\mathbf{Z_{obs}} = \frac{|\bar{x} - m_0|}{\frac{\sigma}{\sqrt{n}}}$
- **Critical value**: $Z_{1-\alpha/2}$ from the standard normal distribution Z.

Confidence interval

$$CI = \bar{x} \pm [\mu_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}]$$

1.4- Conclusion of the test

We accept the hypothesis ($H_0 : \mu = \mu_0$) when:

$T_{obs} < t_{cri}$ or

$Z_{obs} < Z_{cri}$

And we reject this hypothesis ($\mu \neq \mu_0$) when:

$t_{obs} \geq t_{cri}$ or

$Z_{obs} \geq Z_{cri}$

2- Test for comparing two means for independent samples

The test for comparing two means determines whether the means of two populations differ significantly.

2.1- Conditions of application

- The data follows a normal distribution;
- The variances of the samples are equal (homoscedasticity);
- The samples are random and independent.

2.2- Test hypotheses

- **Null hypothesis (H_0)**: $\mu_1=\mu_2$ (the means of the two populations are equal).
- **Alternative hypothesis (H_1)**: $\mu 1 \neq \mu 2$ (the means of the two populations are unequal).

2.3- Test execution

The method used depends on the knowledge of population variances and the sample size.

2.3.1 - Population variance (σ^2) unknown and assumed equal

2.3.1.1 - n_1 or n_2 < 30

In this case, a Student's t-test is used with a pooled variance estimate.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The test statistic is calculated as:

- $t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
- **Critical value:** $t_{1-\alpha/2}$ from Student's t distribution with ($\mathbf{n_1+n_2-2}$) degrees of freedom.

Confidence interval (CI)

$$\text{CI} = (\bar{x}_1 - \bar{x}_2) \pm [\, t_{1\text{-}\alpha/2} \times \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

2.3.1.2 - n_1 and $n_2 \geq 30$

For large samples, even if population variances are unknown, a Z-test can be used by estimating variances from sample variances.

- $Z_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
- **Critical value**: $Z_{1-\alpha/2}$ from the standard normal distribution Z.

Confidence interval (CI):

$$\text{CI} = (\bar{x}_1 - \bar{x}_2) \pm [\, Z_{1\text{-}\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\,]$$

Note: If the variances of the two samples differ, Welch's t-test (an approximation of Student's t-test) is used to adjust the variance and degrees of freedom:

- $t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

Adjusted degrees of freedom (DDL):

- $\text{DDL} = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_{2-1}}}$

2.3.2 - Population variance (σ^2) known

Regardless of sample size, if the population variances (σ_1^2 and σ_2^2) are known, the Z-test is used to compare the two means:

- $Z_{obs} = \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
- **Critical value**: $Z_{1-\alpha/2}$ from the standard normal distribution Z.

Confidence interval (CI): **CI =** $(\bar{x}_1 - \bar{x}_2) \pm [\, Z_{1-\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\,]$

2.4 - Test conclusion

We accept the hypothesis ($H_0 : \mu_1=\mu_2$) when:

$t_{obs} < t_{cri}$ or

$Z_{obs} < Z_{cri}$

And we reject this hypothesis ($\mu_1 \neq \mu_2$) when:

$t_{obs} \geq t_{cri}$ or

$Z_{obs} \geq Z_{cri}$

3 - Test of Comparison of two means for dependent (paired) samples

When samples are dependent (or paired), a test adapted to this situation is used. Paired samples often arise in studies where the same individuals are measured at two different times (e.g., before and after a treatment).

3.1 - Conditions of application

The test application remains subject to the following conditions:

- The data follows a normal distribution;
- The variances of the samples are equal (homoscedasticity);
- The samples are random and dependent.

3.2 - Test hypotheses:

- Null hypothesis (H_0): $\mu_d = 0$ (no difference between the means of the two conditions)
- Alternative hypothesis (H_1): $\mu_d \neq 0$ (significant difference between the means of the two conditions)

3.3 - Test execution

3.3.1 - Variance of differences is unknown

3.3.1.1 - Small sample (n < 30)

- $\mathbf{t_{obs}} = \dfrac{|\bar{d}|}{\frac{S_d}{\sqrt{n}}}$

Where ;

$\bar{d}$ is the mean of the differences between pairs, S_d the standard deviation of the differences and n the number of pairs.

- **Critical value:** $t_{1-\alpha/2}$ from Student's t distribution with $(n - 1)$ degrees of freedom.

3.3.1.2 - Large sample (n ≥ 30)

For large samples, the distribution of differences can be approximated by a normal distribution. A Z-test is used.

- $\mathbf{Z_{obs}} = \dfrac{|\bar{d}|}{\frac{S_d}{\sqrt{n}}}$

- **Critical value**: $Z_{1-\alpha/2}$ from the standard normal distribution Z.

Confidence interval (CI)

$$\mathbf{IC} = \bar{d} \pm \left[Z_{1-\alpha/2} \times \frac{S_d}{\sqrt{n}} \right]$$

3.3.2 - Variance of Differences Known (Rare in Practice)

- $$Z_{obs} = \frac{|\bar{d}|}{\frac{\sigma_d}{\sqrt{n}}}$$
- **Critical value**: $Z_{1-\alpha/2}$ from the standard normal distribution Z.

Confidence interval (CI)

$$\mathbf{IC} = \bar{d} \pm \left[Z_{1-\alpha/2} \times \frac{\sigma_d}{\sqrt{n}} \right]$$

3.4 - Test conclusion

We accept the hypothesis ($H_0 : \mu_d = 0$) when:

$t_{obs} < t_{cri}$ or

$Z_{obs} < Z_{cri}$

And we reject this hypothesis ($\mu_d \neq 0$) when:

$t_{obs} \geq t_{cri}$ or

$Z_{obs} \geq Z_{cri}$

Chapter 3: Statistical methods relating to dispersion

Introduction

This chapter explores the concept of dispersion, a cornerstone of biostatistics that quantifies the spread or variability of data points within a dataset. Variability is inherent in biological and medical data due to factors such as genetic differences, environmental influences, and measurement errors. Understanding dispersion is essential for summarizing data, comparing groups, and making statistical inferences. This chapter covers both theoretical and practical aspects of dispersion measures, with a focus on their applications in biomedical research.

1- Definition

Dispersion refers to the extent to which data points in a dataset differ from each other and from the central tendency (e.g., mean or median).

2- Importance

a. **Accuracy and Precision:** Dispersion measures provide information about the accuracy and precision of statistical conclusions.
b. **Comparison of Data Sets:** It helps in comparing the spread of two or more distributions.
c. **Understanding Variability:** Analyzing dispersion reveals whether data points are closely clustered or widely spread.
d. **Decision Making:** In practical applications like quality control and risk assessment, understanding dispersion aids decision-making.

3. Types of measures of dispersion: Measures of dispersion can be classified into two broad categories: absolute and relative measures.

3.1- Absolute measures: such as range, interquartile range (IQR), variance, and standard deviation, quantify variability in the same units as the data. These measures are useful for understanding the spread within a single dataset **.**

3.1.1- Range

The range is the simplest measure of dispersion, defined as the difference between the maximum and minimum values in a dataset. While easy to calculate, the range is highly sensitive to outliers and does not provide information about the distribution of values within the dataset.

$$\textbf{Range} = \textbf{Max}\ (X) - \textbf{Min}\ (X)$$

3.1.2- Interquartile range (IQR)

The IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Unlike the range, the IQR is less affected by extreme values and is particularly useful for summarizing skewed data.

$$\mathbf{IQR = Q3 - Q1}$$

3.1.2- Variance and standard deviation

Variance and standard deviation are widely used measures of dispersion that quantify the average distance of data points from the mean. The variance is the mean of the square of all deviations from the mean. Mathematically it is given as:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(Population variance)}$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(Sample variance)}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

N.B : In the case of data grouped into k classes of size n_i (continuous variable grouped into classes or discrete variable), the variance formula is as follows:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} n_i(x_i - \bar{x})^2 \qquad \text{(Population variance)}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} n_i(x_i - \bar{x})^2 \qquad \text{(Sample variance)}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$$

The standard deviation is the square root of the variance:

$$S^2 = \sqrt{S}$$

The standard deviation is expressed in the same units as the data, making it easier to interpret. These measures are particularly useful for datasets that follow a normal distribution, as they provide a comprehensive summary of the spread.

3.2- Relative measures

Express variability as a proportion or percentage of a central value, allowing comparison across datasets with different scales.

3.2.1- Coefficient of variation (CV)

The coefficient of variation is a relative measure of dispersion that expresses the standard deviation as a percentage of the mean. It is particularly useful for comparing variability across datasets with different units or scales.

$$\boldsymbol{CV = \frac{S}{\bar{x}} \times 100}$$

3.2.2. Index of dispersion (ID)

The index of dispersion is a measure of variability specifically designed for count data. It compares the variance to the mean and is particularly useful for analyzing datasets where the data points represent counts.

$$\mathbf{ID} = \frac{\sigma^2}{\mu}$$

4. Graphical representations of dispersion

Graphical representations of dispersion are essential for visually interpreting the spread and variability of data.

4.1- Box plots:

Box Plots are commonly used to visualize the interquartile range (IQR), median, and outliers, making them useful for comparing variability across different groups.

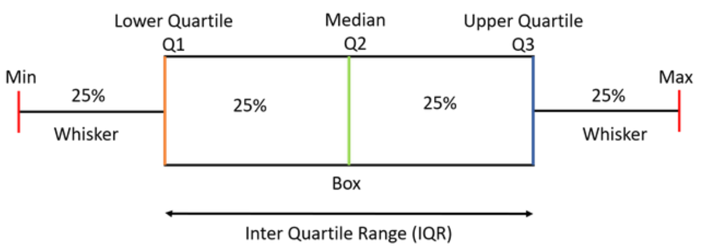

Figure 3. Box Plots

4.2- Error Bars:

Error Bars are employed in bar charts to represent measures such as standard deviation, standard error, or confidence intervals, aiding in the interpretation of data variability.

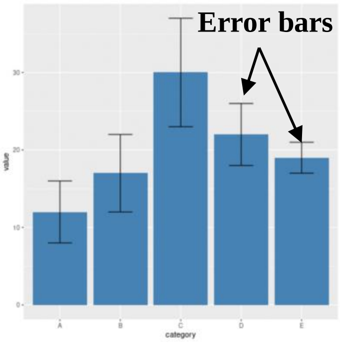

Figure 4. Error bar

4.3- Scatter plots:

Scatter Plots are utilized to display the spread of bivariate data and identify relationships between variables. These graphical methods facilitate a clearer understanding of dispersion patterns and make data analysis more intuitive.

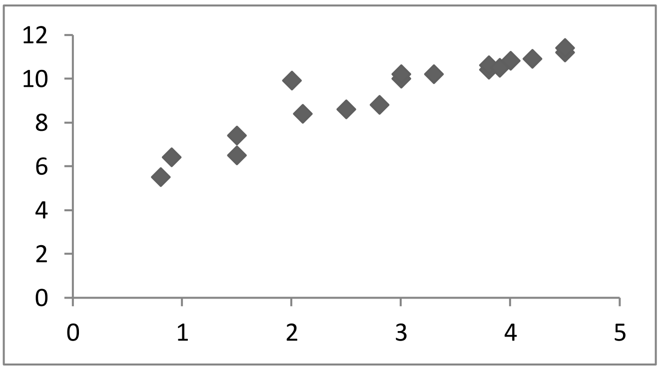

Figure 5. Scatterplots

Chapter 4: Analysis of variance with 1 and 2 classification criteria

Introduction

Analysis of variance (ANOVA) is a statistical method used to test differences between two or more groups based on sample data. It helps determine if the means of different groups are significantly different from each other. There are two main types of ANOVA: One-Way ANOVA and Two-Way ANOVA. These tests are useful in understanding the impact of independent variables (factors) on a dependent variable.

1- One-way ANOVA

1.1- Objective: One-Way ANOVA is used to compare the means of three or more groups based on one independent variable (factor).

1.2- Application conditions

Independence of observations: The samples must be independent of each other.
Normality: The data within each group should be approximately normally distributed.
Homogeneity of variance: The variance within each group should be approximately equal.

1.3- Hypothesis

Null hypothesis (H_0): $\mu_1=\mu_2 = \mu_3 \ldots = \mu_k$ (***k*:** number of groups).
The means of all groups are equal (no significant difference between group means).

Alternative hypothesis (H_1): At least one group mean is different from the others.

1.4- Test execution
1.4.1- Calculation of sum of squares (SS)
1.4.1.1- Total sum of squares (SST)

Measures the total variability in the data.

$$SST = \sum_{i=1}^{N}(y_i - \bar{y})^2 \qquad or$$

$$SST= \sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^{N} y_i)^2}{N}$$

Where:

N : is the total number of observations.
$\bar{y}$ **:** is the grand mean of all observations.

1.4.1.2- Residual sum of squares (SSR)

Measures variation within each group (variation around each group's mean).

$$\boldsymbol{SSR} = \sum_{j=1}^{k} \sum_{i=1}^{nj} \left(\boldsymbol{y_i} - \overline{\boldsymbol{y}}_{\boldsymbol{j}}\right)^2$$

where:

k: is the number of groups.
$\boldsymbol{n_j}$ is the number of observations in group ***j.***
$\overline{\boldsymbol{y}}_{\boldsymbol{j}}$: is the mean of group j

1.4.1.2- Factorial sum of squares (SSF)

$$\boldsymbol{SSF} = \sum_{j=1}^{k} \boldsymbol{n_j}\left(\overline{\boldsymbol{y}}_{\boldsymbol{j}} - \overline{\boldsymbol{y}}\right)^2 \boldsymbol{or}$$

$$\boldsymbol{SSF = SST - SSR}$$

where:

k: is the number of groups.
$\boldsymbol{n_j}$ is the number of observations in group ***j.***
$\overline{\boldsymbol{y}}_{\boldsymbol{j}}$: is the mean of group ***j.***
$\overline{\boldsymbol{y}}$ **:** is the grand mean of all observations.

1.4.2- Degrees of freedom (df)

- **$Df_{factorial}$ (for Between Groups)** = k - 1, where ***k*** is the number of groups.
- **$df_{residual}$ (for Within Groups)** = N - k, where ***N*** is the total number of observations.
- **df_{total}** = N - 1.

1.4.3- Mean Square (MS)

1.4.3.1- Factorial mean square (MSF)

$$\mathbf{MSF} = \frac{\boldsymbol{SSF}}{\mathbf{df_{Factorial}}} = \frac{\boldsymbol{SSF}}{\mathbf{k-1}}$$

1.4.3.2- Residual mean square (MSR)

$$\mathbf{MSF} = \frac{\boldsymbol{SSR}}{\mathbf{df_{residual}}} = \frac{\boldsymbol{SSR}}{\mathbf{N-k}}$$

1.4.4- F-statistic calculation

$$F_{cal} = \frac{SSF}{SSR}$$

1.4.5- F-critical: (F-critical : $F_{1-\alpha;\,(k-1),(N-k)}$)

The F-statistic follows Fisher- Snedecor's F-distribution.

1.4.6- One-way ANOVA table

Source of variation	Df	Sum of squares (SS)	Mean square (MS)	F_{cal}	$F_{critical}$
Factorial	k-1	SSF	MSF	$\frac{MSF}{MSR}$	$F_{1-\alpha\,;\,(k-1)\,(N-k)}$
Residual	N-k	SSR	MSR		
Total	N-1	SST			

1.5- Conclusion of the test:

Compare the F-statistic with the critical value from the F-distribution table based on the significance level (α) and the degrees of freedom.

- If F-Statistic > F-critical, reject the null hypothesis (there is a significant difference between group means).
- F-Statistic ≤ F-critical, accept the null hypothesis (there is no significant difference between group means).

2- Two-way ANOVA

2.1- Objective: Two-Way ANOVA is used to analyze the effect of two independent variables (factors) on a dependent variable, including their interaction effect.

2.2- Application conditions

a. Independence of observations.
b. Normality within each group.
c. Homogeneity of variance.

2.3- Null Hypotheses: there are three null hypotheses:

Null Hypothesis for Factor A (H_0A): The means of Factor A are equal across all levels of Factor B.
Null Hypothesis for Factor B (H_0B): The means of Factor B are equal across all levels of Factor A.
Null Hypothesis for Interaction (H_0AB): There is no interaction between Factors A and B.

2.4- Test execution

2.4.1- Calculate the sum of squares

2.4.1.1- Total Sum of squares (SST): Measures the total variation in the data.

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}\right)^2$$

Where:

Y_{ijk} : is the k-th observation at level iii of factor A and level j of Factor B.
$\bar{y}$ **:** is the grand mean of all observations.
a: is the number of levels of factor A.
b : is the number of levels of factor B.
n: is the number of observations per cell (assuming a balanced design).

2.4.1.2- Sum of squares for factor A (SSA): Measures variation between levels of Factor A.

$$SSA = bn\sum_{i=1}^{a}(y_i - \bar{y})^2$$

2.4.1.3- Sum of squares for factor B (SSB): Measures variation between levels of Factor B.

$$SSB = an\sum_{j=1}^{b}\left(y_j - \bar{y}\right)^2$$

2.4.1.4- Sum of squares for interaction (SSAB): Measures the interaction between Factor A and Factor B.

$$SSAB = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}\right)^2$$

2.4.1.5- Residual sum of squares (SSR): Measures the variation not explained by the factors.

$$SSR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{ij}\right)^2$$

2.4.2-Degrees of freedom (df)

- df for factor A = a - 1, where a is the number of levels of Factor A.

- df for factor B = b - 1, where b is the number of levels of Factor B.

- df for interaction = (a - 1)(b - 1).

- df for residual = ab (n - 1), where n is the number of observations per group.

2.4.3- Mean squares (MS)

2.4.3.1- MSA (mean square for factor A) = SSA / df for Factor A.

2.4.3.2- MSB (mean square for factor B) = SSB / df for Factor B.

2.4.3.3- MSAB (mean square for interaction) = SSAB / df for Interaction.

2.4.3.4- MSR (residual mean square) = SSR / df for residual.

2.4.5- F-Statistics:

F_A (Factor A) = MSA / MSR.

F_B (Factor B) = MSB / MSR.

F_{AB} (Interaction AB) = MSAB / MSR.

2.4.6- Two-way ANOVA table

Source of variation	df	Sum of squares (SS)	Mean square (MS)	F_{cal}	$F_{critical}$
Factor A	a-1	*SSA*	MSA	F_A	$F_{1-\alpha\,;\,(a-1)\,;\,ab\,(n-1)}$
Factor B	b-1	*SSA*	MSB	F_B	$F_{1-\alpha\,;\,(b-1)\,;\,ab\,(n-1)}$
Interaction	(a-1)(b-1)	*SSAB*	MS*AB*	F_{AB}	$F_{1-\alpha\,;\,(a-1)(b-1)\,;\,ab\,(n-1)}$
Residual	ab (n-1)	*SSR*	MSR		
Total		*SST*			

2.5- Conclusion of the test

Compare the F-statistics with critical values from the F-distribution table to determine whether to reject or accept to reject the null hypotheses for each factor and the interaction.

Chapter 5: Dendrometry

Introduction

Dendrometry refers to the process of measuring the diameter of trees. By extension, this term also encompasses the operation of measuring certain quantifiable physical characteristics of trees through various means: diameter, height, volume, shape, age, and bark thickness.

1- Tree size

The size of a tree can be expressed using three measurements:

- Diameter,
- Circumference,
- Basal area.

Diameter and circumference are typically measured at "breast height," which is 1.3 meters above the ground. The basal area corresponds to the cross-sectional area of the tree at 1.3 meters.

1.1- Diameter measurement

1.1.1- Measuring instruments

The diameter of felled or standing trees can be measured using various tools:

a- Forest caliper

The caliper must be held perpendicular to the axis of the tree, with the fixed arm and the rule pressed against the trunk. The sliding arm should then glide against the trunk. The sliding arm should move smoothly while remaining strictly perpendicular to the rule.

Figure 6 . Forest Caliper

b- Finnish Caliper:

This caliper has two fixed arms, one straight and the other curved (parabolic), with bicolored centimeter graduations. The device can be attached to extendable rods, allowing diameter measurements up to about 8 to 10 meters above the ground (or more if binoculars are used for reading). It is held perpendicular to the tree trunk, and the reading is taken along the line of sight parallel to the straight arm of the caliper and tangent to the tree.

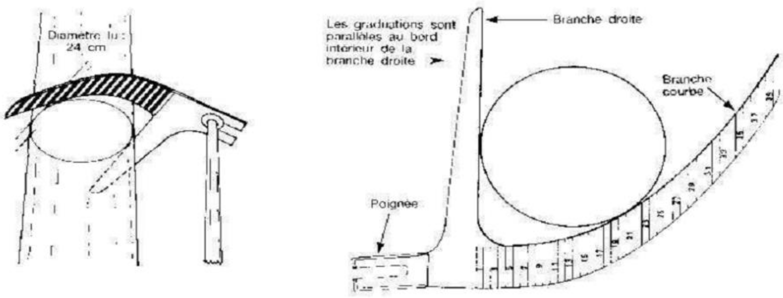

Figure 7. Finnish caliper

a. **Pentaprism**

This tool also allows for diameter measurement at various heights of a standing tree. It consists of optical prisms arranged so that the image of the right side of the trunk provided by a movable prism coincides with the image of the left side of the trunk provided by a fixed prism. The reading is taken on a graduated rule, aligned with a cursor linked to the movement of the movable prism.

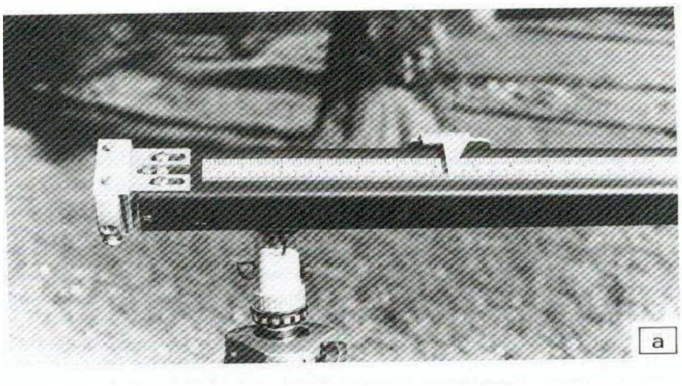

Figure 8. Finnish caliper

b. **Bitterlich relascope**

This device, designed by the Austrian forester Bitterlich, allows for the following main measurements:

- Diameters at various heights ;
- Heights ;
- Form factor ;
- Basal area of a stand;
- Slopes;
- It also enables the deduction of a tree's volume.

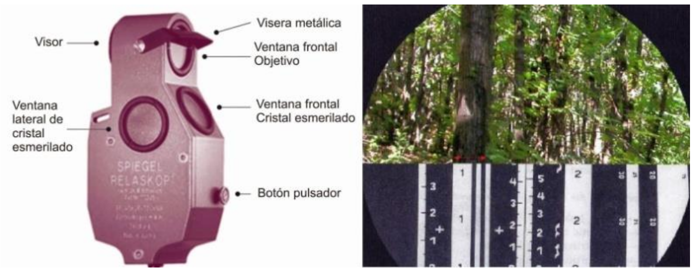

Figure 9. Bitterlich Relascope

1.2- Circumference measurement

Circumference is typically measured using a tape, preferably non-stretchable, made of metal or, better yet, fiberglass. This tape (1.5 m or 3 m) allows for measurements at all levels, with the measurement at breast height considered to be at 1.3 meters. For measuring the circumference at various levels of felled trees, especially those on cutting floors, the use of a tape is not always possible; in such cases, a merchant's string (a curved metal needle attached to a string, the length of which is measured after encircling the tree) is commonly used.

1.3- Basal area determination:

The basal area of a tree is the cross-sectional area of the tree at breast height (1.3 meters).

2- Tree height

2.1- Methods and measuring instruments

We will limit ourselves here to studying the main methods of measuring tree height. We will classify them into two categories:

- Simple methods;
- Dendrometers.

2.1.1- Simple methods:

a. The Biltmore stick:

The "Biltmore Stick," very simple to use, is a direct application of the relationships between similar triangles (Figure 10). This method can be implemented using two sticks of equal length, one held at eye level and directed parallel to the ground, the other held at the end of the first and vertically or parallel to the tree. The operator moves closer to or further from the tree to be measured so that they can simultaneously see the base of the tree (ground level) by aiming at the base of the vertical stick and the top of the tree by aiming at the upper end of the same stick. The height of the tree then corresponds to the distance the operator is from the tree. Indeed, according to Figure below: SP= OA= h

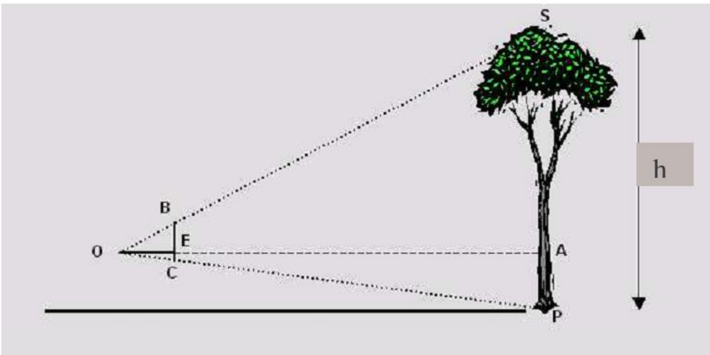

Figure 10. principle of the biltmore stick

2.1.2- Dendrometers:

Among these dendrometers, most of which require the observer to be at a predetermined distance from the tree, it is worth mentioning the BLUME-LEISS, the SUUNTO, the Bitterlich Relascope, and the electronic VERTEX dendrometer.

Figure 11. Dendrometer BLUME-LEISS

3- Tree shape

The dendrometric expressions commonly used to characterize the general shape of a tree (trunk) are:

- Taper coefficient (k),
- Reduction coefficient (r),
- Form factor (f).

3.1- Taper coefficient (k):

This coefficient expresses the ratio between the diameter (or circumference) at mid-height of the stem and the diameter (or circumference) measured at breast height:

$$k = \frac{d_{h/2}}{d_{1.3h}}$$

3.2- Reduction coefficient (r)

The reduction coefficient is the ratio between the difference in size at breast height and at mid-height on one hand, and the size at breast height on the other. It is therefore the complement to 1 of the taper coefficient:

$$\boldsymbol{r = \frac{d_{1.3-}\ d_{h/2}}{d_{1.3}}}$$

3.3- Form factor (f)

The form factor corresponds to the ratio of the actual volume of the tree to the volume of a cylinder with a base equal to the cross-sectional area at 1.3 meters and a length equal to the height h of the tree (at the considered cut):

$$\boldsymbol{f = \frac{v}{g_{1.3} \times h}}$$

Bibliographic references

1. **Chaabane, D., & Zitouni, M.** (2010). *Traité de biostatistique*. Office des publications universitaires.

2. **Ghoraf, N**. *Cours en probabilité et statistique pour 1ère année*. Université Larbi Ben M'hidi Oum El Bouaghi.

3. **Massenet, J. Y.** (2005). *Grosseur des arbres*. Lycée forestier – Château de Mesnières – 76270 Mesnières-en-Bray.

4. **Mezedjri, L**. *Cours en biostatistique*. Université de Skikda.

5. **Mouchiroud, D**. *Mathématiques : Outils pour la biologie – Deug SV1 – UCBL. Chapitre 5, Statistique descriptive*.

6. **Valleron, A.-J.** (2007). *Biostatistique : De la biologie à la clinique*. Flammarion Médecine-Sciences.